和歌の埋め込みに基づく本歌取りの推定

小川隼斗　堀尾海斗　河原大輔
早稲田大学理工学術院
{cookie3120@ruri., kakakakakakaito@akane., dkw@}waseda.jp

概要

本研究では, 和歌に特化した埋め込みモデルとそれを応用した本歌推定モデルを構築する. はじめに事前学習済みの言語モデルを対照学習し和歌に特化した埋め込みモデルを構築する. 次にこのモデルから得られる埋め込みベクトルおよびそれから抽出した特徴量を用いて, 本歌取りをしている和歌の本歌を推定する機械学習モデルを学習する. 本歌とその本歌取りとされる歌のペアデータを用いて本歌推定モデルを評価し, 一定の精度で本歌の推定が可能であることを示した.

1　はじめに

本歌取りとは古歌の歌語の一部や詩情を自作に取り入れて, 表現の重層化をはかり, 自らのうたことばとして表現する手法である [1]. 本歌取りの例を図 1 に示す. 本歌取りの条件を平安時代の歌人藤原定家は歌論書『詠歌大概』で次のように述べている [2].

- 時代の近い歌人の歌は取ってはならない
- 取り入れる古歌の言葉は, 二句及び三, 四字以下にすること
- 主題を変えること

本研究は和歌における本歌取りの自動推定を目的とする. 本歌取りの自動推定では和歌間の共通文字列が有力な手がかりとなる. しかし,『詠歌大概』で述べられている 2 つ目の本歌取りの条件より, 本歌と本歌取りをした和歌には最長共通文字列長が短いことも少なくない. そのため文字列ベースの手法では本歌取りとその他の類似和歌の自動識別が困難である. 本研究ではこの課題に対して, 和歌に特化した埋め込みモデルを構築し, その埋め込みベクトルを基に任意の和歌のペアが本歌取りのペアであるかの確率を出力するモデルの構築に取り組む. 本研究は未発見の本歌取りや引き歌の検出を通じた古典文学研究への貢献につながることが期待される.

図 1　本歌取りの例

2　関連研究

2.1　文字列ベースの類似和歌検出手法

山崎ら [3] や竹田ら [4] は文字列の類似度をベースとした類似和歌の検出手法を提案している. これらの手法により本歌取りや, ある特定の詠歌状況下で用いられる表現, 伝来の過程で表現にバリエーションが生じた異伝歌, 枕詞などの表現技法が共通する歌といった類似和歌の検出が可能である. 一方で, これらの研究は和歌の意味については着目しておらず, 文字列はあまり似ていないが類似和歌である和歌のペアの検出が困難である.

2.2　埋め込みベクトルによる引き歌検出手法

近藤 [5] は, 埋め込みベクトルを用いた引き歌の検出手法を提案している. 引き歌とは本歌取りと類似した表現手法であり, 散文表現の地の文や心情を表した文章に, 著名な和歌の一節を引用する技法 [6] である. この手法ではまず「源氏物語」および「古今集」に含まれるテキストを OpenAI の文埋め込みモデルである text-embedding-ada-002 [7] を用いてベクトル空間に埋め込む. 次に,「古今集」の各和歌の埋め込みベクトルに対して「源氏物語」中の文の埋め込みベクトルとのコサイン類似度を計算し, その値が高いものを引き歌の候補としている. さらに, N-gram の文字列一致によるフィルタリングを適用することで, 引き歌の候補中で引き歌として認定できる「源氏物語」中の文の割合が高ま

ることを報告している. また, この手法によりそれまで未発見であった引き歌も確認している. 一方で,text-embedding-ada-002 による古文の埋め込みの精度と, この手法による引き歌の検出精度については検証していない.

3 和歌埋め込みモデルの構築と評価

和歌埋め込みモデルは, 初めに学習用データセットを構築し, そのデータセットで事前学習済みエンコーダモデルを教師なし SimCSE [8] で学習する. 評価は, 百人一首の和歌とその現代語訳のペアを用いて行う.

3.1 学習用データセットの構築

和歌埋め込みモデルの学習に使用するデータセットとして日本語歴史コーパス (CHJ) [9] に収録されている奈良時代から江戸時代までの文学作品を用いる. またこれに加え, 近代短歌データベース [10] に収録されている短歌および青空文庫で公開されている旧字旧仮名の文学作品も用いる. 日本語歴史コーパス (CHJ) からは和歌約 1 万 7 千首を含む約 10 万文 (以下「CHJ データセット」という), 近代短歌データベースからは和歌約 14 万首 (以下「近代短歌データセット」という), 青空文庫からは約 33.5 万文 (以下「青空データセット」という) を取得した.

3.2 和歌埋め込みモデルの構築

教師あり学習を行うためのアノテーションをデータセット全体に行うのは非常にコストがかかる. 本研究では対照学習である教師なし SimCSE を用いて RoBERTa [11] の日本語モデル[1)]をファインチューニングする. 本モデルにテキストを入力する際は形態素解析器 Juman++ [12] でわかち書きを行う. 教師なし SimCSE は, 文を埋め込みベクトルにする際にドロップアウトを用いることで, 2 つの異なる埋め込みベクトルを取得しそれを正例とみなす. これにより教師データのないデータセットで学習が可能となる. 本研究では 3.1 節で構築した各データセット単一での対照学習と, 複合したデータセットでの対照学習について性能を比較する.

3.2.1 単一データセットによる実験

CHJ データセット, 近代短歌データセット, 青空データセットそれぞれ単体で用いて 5 epoch 学習した.

3.2.2 複合データセットによる実験

青空, 近代短歌, CHJ データセットをマージしてシャッフルしたデータセットで 5 epoch 学習した. さらに, 以下に挙げる段階的に学習データを和歌の形式に近づけるカリキュラム学習を行った. カリキュラム学習において, CHJ データセット以外のデータセットを用いた学習は 1 epoch のみであり, その後 CHJ データセットで 5 epoch 学習している.

- 青空データセット →CHJ データセット
- 近代短歌データセット →CHJ データセット
- 青空データセット → 近代短歌データセット →CHJ データセット

3.3 和歌埋め込みモデルの評価

3.3.1 評価手法

学習した和歌埋め込みモデルの性能を定量的に評価するため, 百人一首の和歌全 100 首とその現代語訳の対訳データセットを用いた評価手法を考案する. 百人一首の対訳データセットは web サイト「百人一首の歴史」[2)3)4)]から取得した. 評価は次の手順で行う.

1. 評価対象モデルを用いて各和歌の原文を埋め込みベクトルに変換する.
2. 同様に 100 首の現代語訳それぞれについても同じモデルで埋め込みベクトルに変換する.
3. 各和歌の原文の埋め込みベクトルと 100 首の現代語訳それぞれの埋め込みベクトル間でコサイン類似度を計算する.
4. 各和歌の原文に対して, 最も類似度が高い現代語訳をモデルが予測した対応訳とみなす.
5. 予測対応訳とその和歌の正解対訳が一致していれば正解とし,100 首に対する正解率で評価する.

3.3.2 評価実験

まず単一データセットによる学習の epoch 毎の正解率の推移を図 2 に示す. 単一データセットでの学習でもっとも性能が良かったモデルは CHJ データセットで 1 epoch 学習させたものであり正解率は

1) https://huggingface.co/nlp-waseda/roberta-base-japanese

2) https://hyakunin.stardust31.com/gendaiyaku.html
3) https://hyakunin.stardust31.com/gendaiyaku-itiran.html
4) https://hyakunin.stardust31.com/yaku.html

表 1　OpenAI モデルと和歌埋め込みモデルの正解率比較.

モデル	text-embedding-small	text-embedding-large	text-embedding-ada-002	和歌埋め込みモデル
正解率	0.95	0.91	0.92	0.95

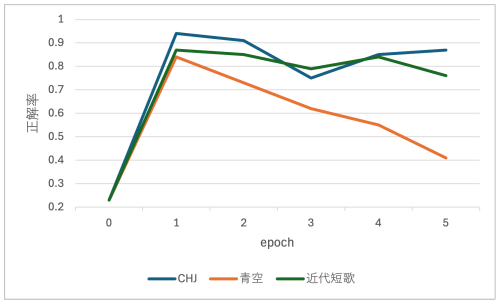

図 2　単一データセットによる学習の epoch 毎の正解率の推移.

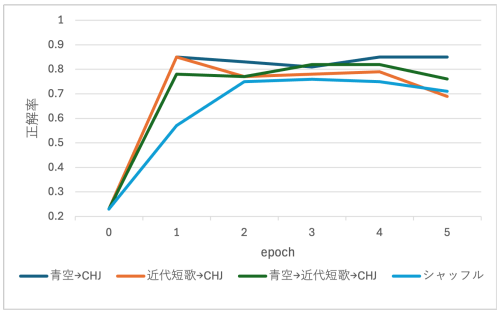

図 3　複合データセットによる学習の epoch 毎の正解率の推移.

0.95 であった. 次に複合データセットによる学習の epoch 毎の正解率の推移を図 3 に示す. 複合データセットでの学習でもっとも性能が良かったモデルは複数あり, 正解率は 0.85 であった. 従って,CHJ データセットで 1 epoch 学習させたモデルがもっとも高い性能を示した. この結果から, CHJ データセットで 1 epoch 学習させたモデルを和歌埋め込みモデルとして採用し, 以降の実験で用いる.

3.3.3　OpenAI モデルとの比較

OpenAI の文埋め込みモデルと本研究で構築した和歌埋め込みモデルを 3.3.2 節で用いた評価手法により比較した. その結果を表 1 に示す. OpenAI モデルでもっとも評価の高かったモデルは text-embedding-3-small であった. その正解率は 0.95 であり和歌埋め込みモデルと同等の性能を示した.

4　本歌推定モデルの構築

4.1　データセット構築

本歌推定モデルの学習および評価のためのデータセットを構築する. まず,「日本うたことば表現辞典 本歌 本節取編」に収録された八代集[5]中の本歌とその本歌取りとされる歌のペア 300 組を人手で収集する. その後, 学習用本歌取りデータセット (200 組) と評価用本歌取りデータセット (100 組) に分割する. また, 負例として八代集中からランダムに組み合わせた和歌のペア 200 組のデータセット (以下「ランダムデータセット」という) を構築する. 最後に「日本うたことば表現辞典 枕詞編 (上・下)」 [13] から枕詞が共通であるが本歌取りではない和歌のペアを 150 組 (以下「共通枕詞データセット」という) を人手で収集する. この際コスト削減のため,10 種の枕詞を選定し各枕詞を含む和歌を 6 首ずつ収集, および 6 首の中から順不同で 2 首の和歌のペアを全通り取り出すことで枕詞 1 種類あたり 15 組のペアを確保している.

4.2　機械学習モデルの構築

和歌埋め込みモデルから得られる埋め込みベクトルを基に本歌推定モデルを構築する. 本歌推定モデルは 2 首の和歌のペアから得られる特徴量を基にその和歌ペアが本歌取りのペアである確率を求める. 本研究で用いる RoBERTa base モデルの埋め込みは 768 次元であり, 和歌ペアの埋め込みベクトルの両方を入力とすると 768×2 次元の入力となる. 学習データの数が少ないため, 学習する機械学習モデルの入力の次元の大きさは抑えたい. そのため入力には埋め込みベクトルを用いず, 代わりに次の 7 つの特徴量を用いる.

- 和歌ペア間のコサイン類似度
- 和歌の各句ごとの埋め込みベクトル同士のコサイン類似度 25 件のうち上位 5 件
- 和歌ペア間の最長共通部分列長

和歌ペア間の最長共通部分列長は, 表記揺れの影響を最小限に抑えるため, 和歌用の形態素解析用辞書

5)　八代集とは 8 つの勅撰和歌集の総称である.

表 2　各手法による本歌推定結果の順位分布.

手法	1 位	2 位	3 位	4 位	5 位
近傍探索	8	9	2	0	1
logistic 回帰	8	9	2	0	1
SVM	7	0	0	0	1
LightGBM	1	6	2	1	2
MLP	9	11	1	2	0
メタモデル	10	5	0	0	0

である和歌 UniDic [14] を用いた MeCab [15] で読みがなを取得し, 濁点, 半濁点を除いた仮名文字に変換し求める. これらの特徴量を用いて logistic 回帰, SVM, LightGBM, MLP およびそれらをブレンディングしたメタモデル (logistic 回帰) を学習する. また,MLP については複数の学習率で学習する. 学習設定の詳細は付録 A に示す.

4.3　実験

4.3.1　評価手法とベースライン

評価手法　日本語歴史コーパス内の八代集に収録されている和歌約 9,600 件 (以下「八代集データセット」という) と評価用本歌取りデータセットを用いて本歌推定の精度を評価する. 評価は次の手順で行う.

1. 評価用本歌取りデータセットの本歌取りとされる歌と八代集データセット全件に対し, 本歌推定手法を適用する.
2. モデルの出力する確率に基づき本歌の可能性が高い順に八代集データセットを並び替える.
3. 次の 2 項目を計算しそのスコアで評価する.
    - Top-5 正解数：並び替えた八代集データセットの上位 5 件に本歌が含まれていた数.
    - MRR (Mean Reciprocal Rank)：正解の本歌が現れた順位の逆数の平均値.

ベースライン　提案手法の比較対象として, 近傍探索を用いた本歌推定をベースラインとする. 評価用本歌取りデータセットの本歌取りとされる歌と八代集データセット全件のベクトル同士のコサイン類似度を計算する. コサイン類似度が大きい順に八代集データセットを並び替えて評価する.

4.3.2　実験結果

機械学習モデルによる本歌推定の Top-5 正解数を表 2 に示す. また, 各モデルの MRR をベースラインの MRR と共に表 3 に示す. logistic 回帰や MLP は比較的良好な結果を示したが,SVM や LightGBM はベースラインのを大きく下回った. Top-5 正解数が最も多かったモデルは MLP（学習率 2e-2,700 エポック）であり 23 件. MRR が最も高かったモデルは MLP（学習率 2e-2,900 エポック）であり 0.172 であり, 近傍探索よりも本歌取りの検出精度が高いと言える. 本モデルが正解の本歌を最も本歌の確率が高いと検出した例を表 4 に示す. その他, モデルの予測で Top-5 に含まれていた本歌取りペアの例を付録 B に示す. MLP の学習率における比較実験の結果のグラフは付録 C に示す.

表 3　各本歌推定手法の MRR.

近傍探索	logistic 回帰	SVM	LightGBM	MLP	メタ
0.149	0.147	0.0793	0.0650	0.172	0.137

表 4　MLP が推定できた本歌取りペアの例.

本歌取りとされる歌	九重の にほひなりせば さくらばな 春知りそむる かひやあらまし
順位	本歌と推定した歌
1 (正解の本歌)	ことしより 春しりそむる 桜花 ちるといふことは ならはざらなん
2	さくら花そこなる影ぞおしまるるしづめる人の春とおもへば
3	さくら花匂ふなごりに大かたの春さへおしくおもほゆるかな

5　おわりに

本研究では、和歌に特化した埋め込みモデルと本歌推定モデルを構築した. さらに, そのモデルの出力する埋め込みベクトルから抽出した特徴量を用いて機械学習モデルを構築した結果, 一定の精度で本歌の推定が可能であることを示した. 本研究には次の課題がある. 和歌埋め込みモデルの入力に用いた Juman++は現代日本語用の形態素解析器であり古文のわかち書きに適していない. 次に学習データの量である. 本歌取りの和歌ペア 300 組を人手で収集したが, より多くのデータを使用できれば精度向上が見込める. また, 和歌埋め込みモデルの評価は古文とその現代語を用いているが, 古文のみで構成された評価データセットを構築することが望ましい. これらの課題を解決することで, モデルの精度向上および古典文学研究への貢献がより一層期待される.

謝辞

本研究にあたり, 近代短歌データベースのデータを提供してくださった村田祐菜氏に心より感謝する. 本研究は JSPS 科研費 JP23K22374 の助成を受けて実施した.

参考文献

[1] 大岡信. 日本うたことば表現辞典 本歌 本節取編. 遊子館, 6 2009.

[2] 『和歌文学大辞典』編集委員会. 和歌文学大辞典. 東京堂出版, 2014.

[3] 山崎真由美, 竹田正幸, 福田智子, 南里一郎. 和歌データベースからの類似歌の自動抽出. 人文科学とコンピュータ, Vol. 40, No. 8, pp. 40–8, 1998.

[4] 竹田正幸, 福田智子, 南里一郎, 山崎真由美, 玉利公一. 和歌データからの類似歌発見. 統計数理, Vol. 48, No. 2, pp. 289–310, 2000.

[5] 近藤泰弘. 『源氏物語』の引き歌をベクトル検索によって検出する方法. **The 38th Annual Conference of the Japanese Society for Artificial Intelligence**, 2024.

[6] 西沢正史. 古典文学を読むための用語辞典. 東京堂出版, 2002.

[7] OpenAI. New and improved embedding model, 2022. Accessed: 2024-10-26.

[8] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[9] 国立国語研究所. 日本語歴史コーパス. version 2024.4, https://clrd.ninjal.ac.jp/chj/.

[10] 村田祐菜. 近代短歌データベース. https://kindaitankadatabase.com/.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv, 2019. abs/1907.11692.

[12] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A morphological analysis toolkit for scriptio continua. In Eduardo Blanco and Wei Lu, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 54–59, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[13] 大岡信. 日本うたことば表現辞典 枕詞編 (上・下). 遊子館, 2007.

[14] 国立国語研究所. 和歌 UniDic, 2023. Accessed: 2024-10-30.

[15] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

A　本歌推定モデルの詳細

A.1　MLP の詳細

- 入力層: 次元数 7
- 隠れ層 1: 全結合層（7 → 32）, 活性化関数: ReLU, ドロップアウト率: 0.2
- 隠れ層 2: 全結合層（32 → 16）, 活性化関数: ReLU, ドロップアウト率: 0.2
- 出力層: 全結合層（16 → 1）, 活性化関数: シグモイド

A.2　メタモデルの詳細

入力特徴量

- 1 層目の予測確率
  - Logistic 回帰の予測確率
  - SVM の予測確率
  - LightGBM の予測確率
  - MLP の予測確率

データ分割設定

- 1 層目の学習に利用した学習データの割合: 0.6
- メタモデルの学習に利用した学習データの割合: 0.4

A.3　各モデルの学習設定

Logistic 回帰

- Maximum iterations: 1000
- Regularization: L2

SVM

- Kernel function: RBF
- Probability estimates: enabled
- Regularization parameter (C): 1.0

LightGBM

- Objective function: binary
- Evaluation metric: binary_error
- Number of boosting rounds: 100

MLP

- Batch size: 8
- Weight decay: 1×10^{-3}
- Optimizer: Adam
- Loss function: Binary Cross Entropy
- Input preprocessing: StandardScaler

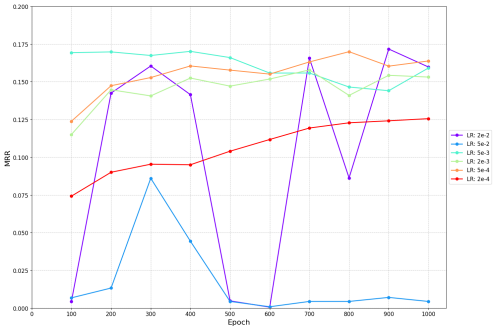

図 4　MLP の学習率別 MRR の推移

メタモデル

- Maximum iterations: 1000
- Regularization: L2
- Tolerance for stopping criteria: 1×10^{-4}

B　本歌推定モデルで推定できた本歌取りペアの例

MLP (学習率 = 2e-2, 700epoch) で推定できた本歌取りペアの例を予測されたスコアの八代集内の順位とともに表 5 に示す.

順位	本歌	本歌取りとされる歌
1	あだなりと 名にこそたてれ 桜花 としにまれなる 人もまちけり	嵐吹く 花の梢は あだなりと 名にこそたてれ 花の白雲
2	けふこずは あすは雪とぞ 降なましきえずは有とも 花とみまし	さくら色の 庭の春風 あともなし 訪はばぞ人の 雪とだにみん
3	山たかみ 人もすさめぬ 桜花 いたくなわびそ 我見はやさむ	春くれど 人もすさめぬ 山桜 風のたよりに 我のみぞとふ
4	花の色は うつりにけりな いたづらに我が身世にふる ながめせしまに	袖の露も あらぬ色にぞ 消えかへる うつれば変る なげきせしまに

表 5　推定できた本歌取りペアの例.

C　MLP の学習率別 MRR の推移

MLP の学習率別 MRR の推移を図 4 に示す.